Cuestionario de teoría 2

Antonio Álvarez Caballero
analca3@correo.ugr.es

13 de mayo de 2016

1. Cuestiones

Cuestión 1. Sean x e y dos vectores de observaciones de tamaño N. Sea

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

la covarianza de dichos vectores, donde $\bar{z}$ representa el valor medio de los elementos de z. Considere ahora una matriz X cuyas columnas representan vectores de observaciones. La matriz de covarianzas asociada a la matriz X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Defina la expresión matricial que expresa la matriz cov(X) en función de la matriz X.

Solución. ss

Cuestión 2. Considerar la matriz hat definida en regresión, $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, donde X es una matriz $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es invertible.

1. Mostrar que H es simétrica
2. Mostrar que $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero K

Solución. Resolvamos ambas partes:

1. Para ver que H es simétrica, no hay más que ver que es igual a su traspuesta. Hay que tener en cuenta que la traspuesta cambia el orden del producto matricial, además de que sabemos que $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es simétrica.

$$H^T = \left(\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\right)^T = \mathrm{X}^{\mathrm{T}^{\mathrm{T}}}\left(\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\right)^T = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-\mathrm{T}}\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} = H$$

2. Para ver que $H^K = K\ \forall K \in \mathbb{N}$, lo haremos por inducción.
<u>$K = 2$</u>: Es claro que $H^2 = H$:

$$H^2 = X\underbrace{(X^TX)^{-1}X^TX}_{I}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

<u>$K \stackrel{?}{\Rightarrow} K+1$</u>: Suponiendo que es cierto para K:

$$H^{K+1} = H \Leftrightarrow \underbrace{H^K}_{Inducción} H = H \Leftrightarrow HH = H \Leftrightarrow \underbrace{H^2}_{K=2} = H$$

Lo cual ya hemos demostrado. Así ya lo hemos probado $\forall K \in \mathbb{N}$

Cuestión 3. Resolver el siguiente problema: Encontrar el punto (x_0, y_0) sobre la línea $ax + by + d = 0$ que este más cerca del punto (x_1, y_1).

Solución. d

Cuestión 4. Consideremos el problema de optimización lineal con restricciones definido por

$$\begin{gathered}\mathrm{Min}_z \mathrm{c}^{\mathrm{T}}\mathrm{z} \\ \text{Sujeto a } \mathrm{Az} \leq \mathrm{b}\end{gathered}$$

donde c y b son vectores y A es una matriz.

1. Para un conjunto de datos linealmente separable mostrar que para algún w se debe de verificar la condición $\mathrm{y_n w^T x_n} > 0$ para todo $(\mathrm{x_n}, \mathrm{y_n})$ del conjunto.
2. Formular un problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, identifique quienes son A, **z**, **b** y **c** para este caso.

Solución. d

Cuestión 5. Probar que en el caso general de funciones con ruido se verifica que $\mathbb{E}_{\mathcal{D}}[E_{out}] = \sigma^2 + \texttt{bias} + \texttt{var}$ (ver transparencias de clase).

Solución. cd

Cuestión 6. Consideremos las mismas condiciones generales del enunciado del Ejercicio.2 del apartado de Regresión de la relación de ejercicios.2. Considerar ahora $\sigma = 0{,}1$ y $d = 8$, ¿cual es el más pequeño tamaño muestral que resultará en un valor esperado de E_{in} mayor de 0,008?.

Solución. Conociendo la expresión del valor esperado del E_{in} y acotándolo inferiormente por 0,008:

$$E_{\mathcal{D}}\left[E_{in}(w_{lim})\right] = \sigma^2\left(1 - \frac{d+1}{N}\right) \geq 0{,}008$$

Ahora sólo sustituimos los valores que nos han dado y despejamos el tamaño de la muestra N.

$$0{,}1^2\left(1 - \frac{9}{N}\right) \geq 0{,}008 \Rightarrow \left(1 - \frac{9}{N}\right) \geq 0{,}8 \Rightarrow -\frac{9}{N} \geq -0{,}2 \Rightarrow \frac{9}{0{,}2} \leq N \Rightarrow 45 \leq N$$

Así, el mínimo N para que el valor esperado de E_{in} sea 0,008 es 45.

Cuestión 7. En regresión logística mostrar que

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1 + e^{y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$$

Argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.

Solución. Partimos de la expresión de E_{in} conocida:

$$E_{in}(w) = \frac{1}{N}\sum_{n=0}^{N}\ln\left(1 + e^{-y_i w^T x_i}\right)$$

Ahora sólo calculamos su gradiente con respecto a w, por lo que coincide con su parcial con w.

$$\nabla E_{in}(w) = \frac{\partial}{\partial x}\left(\frac{1}{N}\sum_{n=0}^{N}\ln\left(1+e^{-y_i w^T x_i}\right)\right) = \frac{1}{N}\sum_{n=1}^{N}\frac{-y_n x_n e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N}\frac{-y_n x_n}{1+e^{y_n w^T x_n}}$$

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1+e^{y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$$

Es claro que una solución mal clasificada influye más que una buena. Una solución bien clasificada equivale a $y_n w^T x_n > 0$, luego la exponencial del denominador será mayor que 1. Al dividir por algo más grande, el cociente se hace más pequeño. En el caso de $y_n w^T x_n < 0$, la exponencial será una exponencial negativa, con valores menores a 1. Luego el denominador será más pequeño, y el cociente más grande. Luego la sumatoria será más grande.

Cuestión 8. Definamos el error en un punto (x_n, y_n) por

$$e_n(w) = \max(0, -y_n w^T x_n)$$

Argumentar que el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\nu = 1$.

Solución. asdffg

Cuestión 9. El ruido determinista depende de $\mathcal{H}$, ya que algunos modelos aproximan mejor f que otros.

1. Suponer que $\mathcal{H}$ es fija y que incrementamos la complejidad de f.
2. Suponer que f es fija y decrementamos la complejidad de $\mathcal{H}$

Contestar para ambos escenarios: ¿En general subirá o bajará el ruido determinista? ¿La tendencia a sobrejaustar será mayor o menor? (Ayuda: analizar los detalles que influencian el sobreajuste)

Solución. dfasdf

Cuestión 10. La técnica de regularización de Tikhonov es bastante general al usar la condición

$$w^T \Gamma^{\mathrm{T}} \Gamma w \leq C$$

que define relaciones entre las w_i (La matriz Γ_i se denomina regularizador de Tikhonov)

1. Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$
2. Calcular Γ cuando $(\sum_{q=0}^{Q} w_q)^2 \leq C$

Argumentar si el estudio de los regularizadores de Tikhonov puede hacerse a través de las propiedades algebraicas de las matrices Γ.

Solución. asdf

BONUS. Considerar la matriz $\hat{\mathrm{H}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$. Sea X una matriz $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ invertible. Mostrar que traza$(\hat{\mathrm{H}}) = d+1$, donde traza significa la suma de los elementos de la diagonal principal. (+1 punto)

Solución. Para resolver esto sólo necesitamos utilizar una propiedad básica de álgebra lineal:

$$\text{Traza(AB)} = \text{Traza(BA)}$$

Utilizando esto, dividimos $\hat{H}$ en dos matrices A y B.

$$A = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$$
$$B = \mathrm{X}^{\mathrm{T}}$$

Es claro que $AB = \hat{H}$. Es claro que BA es la matriz identidad. Lo único que hay que tener cuidado es con las dimensiones de esta matriz.

$$\mathrm{X} \in \mathcal{M}_{N\times d+1} \Rightarrow \mathrm{X}^{\mathrm{T}}\mathrm{X} \in \mathcal{M}_{d+1\times d+1}$$

Como una matriz y su inversa tienen la misma dimensión, entonces:

$$BA = \underbrace{X^T X}_{\mathrm{M}} \underbrace{(X^T X)^{-1}}_{\mathrm{M}^{-1}} = I_{d+1}$$

Y la traza de esta matriz es $d+1$. Como coincide con la traza de $\hat{H}$, podemos concluir que

$$\text{traza}(\hat{\mathrm{H}}) = d+1$$